TENSOR: retrieval and analysis of heterogeneous online content for terrorist activity recognition

AKHGAR, Babak <http://orcid.org/0000-0003-3684-6481>, BERTRAND, Piere, CHANANOULI, Christina, DAY, Tony <http://orcid.org/0000-0002-3214-6667>, GIBSON, Helen <http://orcid.org/0000-0002-5242-0950>, KAVALLIEROS, Dimitrios, KOMPASTSIARIS, Ioannis, KYRIAKOU, Eva, LEVENTAKIS, George, LISSARIS, Euthimios, MILLE, Simon, TSIKRIKA, Theodora, VROCHIDIS, Stefanos and WILLIAMSON, Una

Published version

AKHGAR, Babak, BERTRAND, Piere, CHANANOULI, Christina, DAY, Tony, GIBSON, Helen, KAVALLIEROS, Dimitrios, KOMPASTSIARIS, Ioannis, KYRIAKOU, Eva, LEVENTAKIS, George, LISSARIS, Euthimios, MILLE, Simon, TSIKRIKA, Theodora, VROCHIDIS, Stefanos and WILLIAMSON, Una (2017). TENSOR: retrieval and analysis of heterogeneous online content for terrorist activity recognition. In: Proceedings Estonian Academy of Security Sciences, 16 : From Research to Security Union. Estonian Academy of Security Sciences, 33-82.

TENSOR: RETRIEVAL AND ANALYSIS OF HETEROGENEOUS ONLINE CONTENT FOR TERRORIST ACTIVITY RECOGNITION

Babak Akhgar, PhD
CENTRIC (Centre of Excellence in Terrorism, Resilience, Intelligence and Organised Crime Research), Sheffield Hallam University, UK
Director of CENTRIC, Professor of Informatics

Pierre Bertrand
Thales Group, La Défense, France
Data Scientist

Christina Chalanouli, PhD
KEMEA, Greece
Senior Researcher

Tony Day, BSc
CENTRIC, Sheffield Hallam University, UK
Technical Lead

Helen Gibson, PhD
CENTRIC, Sheffield Hallam University, UK
Lecturer in Computing

Dimitrios Kavallieros
KEMEA, Greece
Senior Researcher

Emmanuel Kermitsis
KEMEA, Greece
Senior Researcher

Ioannis Kompatsiaris, PhD
Information Technologies Institute
Centre for Research and Technology Hellas, Greece
Senior Researcher (Researcher A)

Eva Kyriakou
European Organisation for Security, Belgium
Project and Policy Manager

George Leventakis, PhD
KEMEA, Greece
Senior Advisor and Scientific Coordinator in European Projects

Euthimios Lissaris
KEMEA, Greece
Police Captain

Simon Mille, PhD
Universitat Pompeu Fabra, Spain
Researcher

Dimitrios Myttas
KEMEA, Greece
Senior Researcher

Theodora Tsikrika, PhD
Information Technologies Institute
Centre for Research and Technology Hellas, Greece
Postdoctoral Research Fellow

Stefanos Vrochidis, PhD
Information Technologies Institute
Centre for Research and Technology Hellas, Greece
Senior Researcher

Una Williamson
Police Service of Northern Ireland
Head of International Programmes

Keywords: counterterrorism, data mining, dark web, social media

ABSTRACT

The proliferation of terrorist generated content online is a cause for concern as it goes together with the rise of radicalisation and violent extremism. Law enforcement agencies (LEAs) need powerful platforms to help stem the influence of such content. This article showcases the TENSOR project which focusses on the early detection of online terrorist activities, radicalisation and recruitment. Operating under the H2020 Secure Societies Challenge, TENSOR aims to develop a terrorism intelligence platform for increasing the ability of LEAs to identify, gather and analyse terrorism-related online content. The mechanisms to tackle this challenge by bringing together LEAs, industry, research, and legal experts are presented.

1. INTRODUCTION

For most citizens, the Internet is a valuable resource in day-to-day life, but for criminals and terrorists, it provides opportunities to exploit the Web as a tool where they can communicate with affiliates, coordinate action plans, raise funds and introduce new supporters or recruits into their networks. These activities present a significant risk to the citizens of Europe.

TENSOR is an EU project funded under the Secure Societies pillar of the Horizon 2020 programme and aims to develop a platform offering Law Enforcement Agencies (LEAs) fast and reliable planning and prevention functionalities for the early detection of terrorist activities, radicalisation and recruitment. The project aims to develop solutions to mitigate this risk from terrorism and prevent future attacks or crimes from occurring by analysing potential terrorism-related content resulting from extremists' open communications and activity patterns online. To this end, a unified platform will be developed, which will allow for multidimensional content integration from heterogeneous online resources, with a view to gathering large amounts of Surface, Deep, and Dark Web content, applying automatic analysis and summarisation, and presenting the collected intelligence through intuitive interactive interfaces.

Informed by the requirements of LEAs and the challenges they face, the TENSOR platform will include beyond state-of-the-art techniques for searching, crawling, monitoring and gathering multimodal and multilingual Web content with the aim of expanding LEAs current reach and information sources. The techniques developed in TENSOR aim to improve efficiency, performance and effectiveness in finding and gathering this content. Once the TENSOR platform has successfully acquired content, information extraction techniques will be employed such as entity-extraction, image, video and audio recognition, as well as automatic translation. This will allow for the content to be categorised against a custom-developed taxonomy for terrorist-generated content. Categorisation will provide the basis for the TENSOR platform to perform an automated analysis of the content, employing techniques such as clustering and classification, social network analytics and semantic

reasoning. After the automated analysis has been performed, the platform will automatically select the most relevant content and generate summaries and visualisations to be displayed to the end-user LEAs. This is expected to significantly reduce "information overload" on LEAs and contribute to an increase in efficiency and performance in analysing terrorist-generated content online. To this end, most processes are automated in TENSOR, however end-users will have the option of reviewing these processes and re-configuring the system, making sure that the outputs fit with what is required, should there be a need to do so. Moreover, EU data protection regulations will be taken into account during the design and development of the system. Measures will be taken to ensure that the principles of a) data minimisation, b) data quality, c) data limitation, d) data protection, e) data portability, and f) data breach notifications are built into the system.

By delivering these capabilities, it is expected that TENSOR will positively impact upon: a) more efficient and effective prevention of terrorist activities organised and planned online; b) faster detection of novel terrorism and radicalisation trends, terrorist-published content and grassroots terrorist cells; c) reduction of "information overload" on LEAs, by automatically summarising and visualising only the relevant content; d) built-in privacy and data protection; e) industry's understanding of LEA requirements, and therefore a positive impact on the development of future products and Europe's overall industry competitiveness.

This article showcases the TENSOR project by presenting the challenges LEAs face and the methodology applied in TENSOR for extracting their requirements (Section 2), the tools and technologies currently being developed as part of the integrated TENSOR platform which aim to advance the state-of-the-art in acquiring, analysing, summarising and visualising terrorism-related Web content (Section 3), a legal and ethical assessment of the current operational environment in several European countries (Section 4), and the impact TENSOR may have on this domain (Section 5), before concluding (Section 6).

2. TENSOR USE CASES: CHALLENGES, METHODOLOGY, AND USER REQUIREMENTS

TENSOR employs an agile user-centred methodology to inform the development of the platform. This includes close consultation with a number of LEAs and security organisations (such as the Police Service of Northern Ireland (UK), Mossos d'Esquadra (Catalonia), National Crime Agency (UK), West Yorkshire Police (UK), Belgian Federal Police, as well as organisations from Greece and Germany) to develop a comprehensive set of requirements for the platform.

These user requirements were created based on specific use case scenarios in four areas pertinent to terrorism: domestic terrorism, international terrorism, lone actor terrorism and radicalisation, and are based on real life events encountered by the LEA partners in the project. Through these scenarios, it was possible to extract and analyse the challenges of LEAs and define the capabilities needed to effectively overcome these challenges. The user requirements were subsequently distilled from the scenarios in the form of Agile User Stories and were gathered first at a high and then at a detailed level. Analysing each high-level requirement into a subset of lower level requirements led to the identification of corresponding functional and non-functional requirements.

Next, we first describe the challenges faced by LEAs while fighting terrorism on the World Wide Web and then depict the use cases/user-requirements.

2.1 CHALLENGES

Through the four use cases and preliminary requirements gathering, the project was able to recognise a number of key challenges that are particularly significant in the terrorism domain and that the TENSOR platform would need to tackle in order to provide functionality beyond that the LEAs already have access to. This section describes those challenges in the context of what information would be required and how such information may currently be used by terrorists; this is classified into the following major categories:

- **Utilising the Surface, Deep, and Dark Web as tools for coordinating, recruiting, training and planning of terrorist acts**: Terrorist groups (and their supporters) use the Web to recruit and train new members, and organise coordinated attacks. Especially the Deep and Dark Web can provide, due to their anonymous and encrypted structure, a safe environment for coordinating and planning attacks, minimising the chances of detection and arrest. They are also used for training and radicalisation by facilitating the sharing and dissemination of information and knowledge (e.g., hacking instructions, calls to actions, propaganda) to less experienced supporters without revealing the publisher's identity or location. Moreover, Hidden Service Marketplaces (HSMs) that exist on the Dark Web, particularly in TOR and the IP2 may remain unknown to LEAs for some time. The automated monitoring of such websites, marketplaces, forums, and social media, using word/video/image recognition software, will enable LEAs to search and scrape online data in a timely manner.
- **Accessing social networks and closed groups:** Gaining access, monitoring, acquiring evidence, and investigating "closed groups" on social networks and closed forums can be challenging. Investigators must often wait several weeks before requesting access to their administrators, as time is needed for accounts to seem authentic and to develop a realistic backstory. Amplifying individual weak-signals of online activities by grouping them together with other behavioural and contextual factors, of the same and/or other persons, can provide a comprehensive picture allowing authorities to assess the level of a potential threat to society. The need to engage in covert Web investigation to elaborate on their suspicions and build a body of evidence requires LEAs to invest in operation time which may not produce concrete results, but only further circumstantial evidence. Thus, an automated process that is able to predict, exploit, and respond authentically to common interaction requests on social media and Dark Web forums could free up investigators' time and gain access to these closed groups, before being taken over by analysts once specific information is required.
- **Extraction and analysis of multilingual multimedia content:** The Web is not simply text-based, but is composed of multiple different content types (including images, video, and audio) published and posted in different languages, utilising different colloquialisms and

idioms (e.g., arabizi). For standardisation, any extracted content needs to be accurately transcribed, translated, analysed, and categorised into languages preferred by the end users of the TENSOR platform so as to be "interpreted" correctly, and in a timely manner, to assist LEA experts in determining their appropriate Course of Action (CoA).

- **Understanding and identifying terrorists' perspectives:** To prevent terrorism and to successfully tackle the underlying causes of radicalisation, LEAs are required to gain a greater understanding at an early stage of the psychological preparations and perspectives of violent extremists, their religious and ideological beliefs and the consequential societal influences. These beliefs form the cornerstone for the claims and desire to fight for religious causes, as well as providing the foundations for developing extremist and fanatic attitudes, their subsequent aspirations to convert unbelievers, and their drive to advance their efforts toward their perceived moral betterment of society and social values, such as social justice, solidarity and freedom.

2.2 USER REQUIREMENTS DEVELOPMENT

The requirements were developed based on the identification of user groups and user stories. Two likely user groups of the TENSOR platform, as determined by TENSOR partner LEAs, are intelligence officers and operational intelligence analysts. These user groups were assigned to a number of user stories based on what their operational activities would be. Each user story forms part of a greater whole known as an "epic" which encapsulates a larger use case. Furthermore, the stories were also assigned to three categories: ingestion, analysis, and storage. These attributes were combined with the LEAs' requirements of the TENSOR platform. These correlations allow each user story to address a single requirement, assisting the technical partners to understand the outcome described and the situation that this requirement will resolve, as the following figure depicts.

FIGURE 1: AGILE USER STORY EXAMPLE

#	Type	Category	As an...<Actor>	I can...<Activity>	So that...<Effect>
01	Epic	Analysis	Operational Intelligence Analyst	Determine whether my suspect is already known to authorities as a person of interest or involved in known terrorist/ organised crime groups or online communities	I know as much as possible about my suspect and their history
02	Story	Ingestion	Operational Intelligence Analyst	Ingest a list of persons of interest into the TENSOR platform	The platform knows the persons of interest that I am interested in

Next, the tools and technologies being developed in TENSOR for satisfying the distilled user requirements are presented.

3. TOOLS AND TECHNOLOGY IN TENSOR

The TENSOR platform is composed of a number of components which will be integrated into a unified platform (see Section 3.4). These components include the methods for identification and extraction of online terrorist content (Section 3.1), the analysis of this extracted (textual and multimodal) content (Section 3.2), and the summarisation, presentation, and visualisation of the analysed content for consumption by LEAs (Section 3.3).

3.1 TERRORIST-GENERATED CONTENT ACQUISITION, PROCESSING AND INDEXING

The discovery and acquisition of online terrorist-generated content is the foundation of the TENSOR platform and all other components depend on the provision of this content. We consider online terrorist-generated and terrorism-related content to correspond to textual and multimedia information available on the Surface, the Deep and Dark Web.

3.1.1 TENSOR Data Models and Sources

Each individual piece of content is referred to as an *artefact*. Examples of artefacts may include, among other things, documents, articles, videos, blog posts, comments and likes. Each artefact may possess many *attributes*, some that are selected and extracted from within the original meta-data and others that are attached to the artefact throughout the various stages of processing. Attributes describe various aspects of the artefact, such as when and from where it was obtained, its unique identity, and may well reference other artefacts.

Entities are extracted from the actual content of the artefact and represent the *things* within the content, e.g., organisations, locations, objects, etc. TENSOR is currently developing a comprehensive taxonomy and ontology of terrorism-related entities and classes, as well as the indicators

required for extracting them from acquired artefacts. There are specific components being developed which will extract entities from both textual and multimedia content in various supported natural languages and dialects. Such extraction gives TENSOR components their mechanism for understanding and reasoning against terrorism-related content.

Artefacts and their entities will be *linked* together, allowing for a graph-based model to form. It is these links that allow patterns in the data to emerge through the various processing mechanisms that will be built into the TENSOR processing pipeline.

Acquiring artefacts, entities and links will take place through the various types of sources available, both open public and restricted. These sources may be grouped into four distinct tiers when considering both the nature of availability and content privacy; these tiers are:

1. **Tier 1: Open public non-personal data**, including all widely available published content such as traditional news media sources, widely recognised blogs, web feeds (i.e., RSS) and public social media streams from organisational groups. This tier of content can, but most often does not contain sensitive or personal data, and is usually matter of fact information.
2. **Tier 2: Open restricted non-personal data**, including generally publically available Web forums and social media groups, which although often involves data created by potentially identifiable individuals, is often topical information and not personal in nature.
3. **Tier 3: Open public personal data**, including public facing social media profiles and posts which although are often publicly available, the author has not necessarily explicitly intended the content to be made public. This also covers information that is more likely to contain personal discussions, such as social media users sending publicly visible messages or comments to each other.
4. **Tier 4: Open restricted personal data**, including anything that requires both authenticated access and an *insider* profile or avatar that is in some way connected to an individual or group in order to monitor or acquire their data. This is the most invasive tier of content acquisition and should only be used in specific investigative scenarios where the appropriate authorisation is in place.

Other guidance on these tiers and the levels of authorisation required for them may come from sources such as the UK's National Police Chiefs Council (2015) who define levels of open source investigation and research based on the extent to which they are overt or covert investigations. As TENSOR may operate across the EU, it must be mindful of the differing legislations in different countries (see Section 4).

3.1.2 Content Acquisition, Crawling, and Extraction

Extraction of content from the Surface, Deep, and Dark Web poses a range of challenges for the implementation of the Web crawlers and scrapers that are employed to obtain such content. On the Surface Web, although there exists a greater body of research, many recent reviews (e.g., Weninger, 2016) have noted that extraction methods fail to keep up with current Web trends and the dynamic content that is often served to the user. One recommendation is to make use of headless browsers to ensure the page is fully rendered before extraction while another is to consider the evolving standards that are being brought in by HTML5. Content on the Deep Web is often hidden behind logins and captchas that make automated access more complex. Furche et al. (2013), He (2013), and Zhao et al. (2016) have all recently proposed mechanisms such as adaptions to the XPath extraction method, using reverse link searching to identify Deep Web sites in the first place, or using specific extraction methods for obtaining content from 'entity' based sites.

The Dark Web provides further crawling, mining and extraction challenges as site discovery in the first place is often more complex. Furthermore, many of these sites may be 'invite only' or only appear for limited time periods. Bouchard et al. (2014) have already proposed a system for distinguishing between terrorist and non-terrorist sites on the Dark Web, in particular noting that the phraseology used on the two types of sites differs massively. Chen (2011) offered a number of suggestions for mining the Dark Web, while Zhou et al. (2005) introduced a knowledge management portal for the storage and retrieval of information relating to terrorist groups on the dark web.

Even within these systems for accessing information across the different layers of the Web, there also remains a consideration around how much

autonomy the TENSOR platform should have when accessing this content. Too much autonomy and there is a risk of the system being accused on conducting surveillance, while too little autonomy will not reduce the burden on intelligence analysts' workload and the platform will not be used to its full potential.

3.1.3 Operational Mechanisms

The mechanisms TENSOR aims to use for acquiring content can be broken down into two categories, active and passive. Active content acquisition covers most uses of the TENSOR content acquisition tools. It involves actively making requests against online services for specific types of content via searches and crawling, both of which can leak information to the service about the tool's intentions. Passive acquisition on the other hand attempts to take a more hands-off approach, by exploiting technologies for monitoring purposes. These technologies include RSS feeds, social media streaming sources, newsletter subscriptions and mailing lists. Passive approaches will be employed as much as possible due to their less revealing approach. However, the need for active mechanisms emerges during targeted investigations. Nonetheless the salient point to keep in mind when conducting such research is the mantra of necessary, proportionate and justified (Association of the Chiefs of Police Officers (ACPO), 2013). TENSOR will employ mechanisms to resist the function creep that pervades in many social media based research tools for law enforcement (Trottier, 2013) and the tendency to keep data long beyond its usefulness.

During the retrieval stage, all content that is acquired will be given a unique identity within the TENSOR platform. Secure Hashing Algorithm (SHA) (see Section 3.4.4) will be employed to provide a verifiable identifier for the content to enable the detection of duplicates as well as to protect content from tampering. On top of hashing, Digital Signature Algorithm (DSA) (see Section 3.4.4) will add tamper proofing to the verifiable identifier (or hash) allowing downstream components and subjects to verify the integrity of the content.

In subsequent stages, TENSOR aims to effectively filter and anonymise all acquired textual and multimedia content. Anything not meeting the

minimum required attributes to be considered terrorist-generated or related will be removed immediately upon detection. Further processing will take place in the TENSOR processing pipeline in order to extract entities and links between the discovered and acquired artefacts; some of this extraction and classification will lead to further cyclical searches to discover more relevant content.

TENSOR aims to store and manage all artefacts, entities, and links in a generic and extensible manner. These are the main types of data within the data acquisition process before additional processing takes over. The use of a generic approach enables simplified indexing of content within underlying database technologies. For example, entities can capture various types of classes simultaneously such as locations and categories, which can be indexed together. This does result in fewer, longer indexes, but advantageously provides the capability to deal with entity types that were previously unknown.

3.2 MULTI-MODAL CONTENT ANALYSIS

The analysis and correlation of information extracted from multimodal content aims to ultimately provide LEAs with threat assessment and early warning capabilities, by uncovering the structure underlying the terrorism-related information and data through clustering, classification, community detection and key player identification in social networks, information source quality assessments, multimedia forensics as well as semantic reasoning and enrichment.

3.2.1 Clustering

Clustering aims to group together multimodal objects about similar topics so as to reduce information overload and increase corroboration through aggregation of multiple sources containing the same information. To this end, TENSOR first applies Formal Concept Analysis (FCA) (Ganter & Wille, 1998) using InClose (Andrews, 2011), a deterministic method of deriving a set of hierarchical clusters, each containing a set of instances (multimodal objects) that share a number of common

attributes, such as the terrorism-related entities identified in the TENSOR taxonomy (including people, objects, locations and events), categories, sources, and extracted keywords. The further down the hierarchy one travels, the more specific (more attributes, fewer instances) each cluster becomes. Both instances and attributes can appear in multiple clusters.

Moreover, clustering in TENSOR also relies on methods applied on a graph of the multimodal objects, where nodes usually represent tuples of multiple modalities (e.g. text-image pairs) and links between any two nodes are assigned in an unsupervised or semi-supervised way (Petkos et al., 2017). Community detection on this graph provides densely connected patterns of mutually related objects, resulting in communities of objects that share similar topics. Extracting the correct number of topics is equivalent to the estimation of the correct number of clusters; these are typically not known a priori. To estimate this number, TENSOR relies on multiple realisations of approaches such as DBSCAN*-Martingale (Gialampoukidis et al., 2016b); such methods are robust to noise (i.e. can deal with data not belonging to any topic) and are also able to scale efficiently.

Experiments performed to evaluate the proposed DBSCAN*-Martingale against well-established and parameter-free community detection algorithms were based on four realistic benchmark networks developed by Lancichinetti et al. (2008). The results indicate improvements in the effectiveness of the proposed DBSCAN*-Martingale community detection algorithm in terms of the Normalised Mutual Information (Danon et al., 2005) and RAND (Rand, 1971) metrics. In particular, the most significant differences to the other approaches for both evaluation metrics are observed for the smallest dataset where DBSCAN*-Martingale indicates improvements ranging from 12% to 35% in terms of NMI and from 5.6% to 8.8% in terms of RAND. In the larger datasets, DBSCAN*-Martingale still performs better than all the other approaches, but the differences in the effectiveness are smaller, particularly for the RAND evaluation metric. The second best performing community detection approach is Walktrap (Pons & Latapy, 2006), with the exception of NMI for the smaller datasets, where the Fast Greedy (Clauset et al., 2004) and the Louvain (Blondel et al., 2008) methods perform second best.

3.2.2 Classification

Classification aims to automatically assign the multimodal objects to specific categories, e.g. regarding the level of radicalisation exhibited by a document consisting of multiple modalities (such as a Web page or social media post) using machine learning and deep learning techniques that exploit the rich information from the different modalities (e.g. text and images) and the inter-connections among them. TENSOR first employs Recurrent Neural Networks (RNNs) to build a text-based model that is learnt based on a set of documents annotated with the specific categories of interest. Given a new document, the model projects it into the produced latent vector space and classifies it to an appropriate category. Regarding images, the same methodology is applied with the only difference that the first layer uses Convolutional Neural Networks (CNNs) instead of RNNs. During these two classifications, latent vectors are extracted for representing respectively the two modalities (i.e. texts and images) into similar spaces so as to merge them. Finally, a third model is learnt to perform the classification by fusing together the two modalities; in this last case, the main challenge is to deal with documents without images and thus make the third model adaptive to this missing input. Preliminary experiments indicate promising results for the individual modalities using RNNs and CNNs respectively, while further research is needed for their combination.

3.2.3 Social Network Analysis

Social Network Analysis aims to detect communities of users (e.g., user accounts on forums/social media platforms) engaging in suspicious terrorism-related communications and also identify the most important and influential actors with a key role in the connectivity of the social network and thus the dissemination of information. For instance, Twitter has been extensively used for promoting and spreading terrorism-related propaganda due to its nature that permits the inexpensive communication of multimodal messages (tweets) to users worldwide; to this end, a top-down approach is often used, with a core group of members spreading a terrorist group's messages, which are then re-shared by other affiliated accounts. For both LEAs and the administrators of social media networking platforms, it is of vital significance to prevent terrorist

groups from spreading their propaganda (to the extent possible), by shutting down accounts who are found to play a central role in this information exchange.

To this end, TENSOR employs centrality measures and in particular entropy-based centrality measures, such as the Mapping Entropy (ME) and the Mapping Entropy Betweenness (MEB) (Gialampoukidis et al., 2016a). Intuitively, one may think of a random walker on the network, standing at a node who picks his/her next step with a probability equivalent to the degree centrality (in the case of ME) and equivalent to the betweenness centrality (in the case of MEB) and is summed over all neighbours of a node. These two measures consider the information that is communicated through nodes who act as a hub (bridge), i.e. those with high values of degree (betweenness) centrality between any two members. In particular, the MEB centrality considers the betweenness centrality of a node and also exploits local information from its neighbourhood; hence, high MEB values indicate that a particular node can act as a bridge for disseminating information, even if their degree centrality is low. In parallel to the key-player identification, a community detection algorithm is used to divide the network into groups of users (communities). The top-ranked key-player is used to enrich the retrieved results, which is achieved by searching for the community the key-player belongs to (Gialampoukidis et al., 2017).

The proposed centrality measure was evaluated in a network formed by user mentions in terrorism-related Twitter accounts, which were retrieved using a set of five Arabic keywords related to terrorist propaganda. As ground-truth, account suspension information from Twitter was used, which marks user accounts as suspended, given that the suspension process is applied when an account violates Twitter rules by exhibiting abusive behaviour, including posting content related to violent threats and hate speech. The top-100 user accounts identified as key players were examined to determine whether they are suspended, active or no longer exist (i.e., accounts which have been temporarily or permanently deactivated). The results indicate that the entropy-based centralities ME and MEB are able to retrieve the first suspended user at position 16, while PageRank follows at position 19. Other centrality and popularity measures, such as closeness, eigenvector and number of followers do not find any suspended users in the top-100 positions of their retrieved

users. We observe that the network is very spread with many bridges and a diameter equal to 27, so key players are expected to be positioned in between many pairs of nodes in the network, exploiting also their neighbourhood's high betweenness centrality.

3.2.4 Information Source Quality Assessment

Information source quality assessment employs a multi-dimensional viewpoint to interpret the notion of "quality", e.g. in terms of reliability, credibility, relevance, precision, etc. This is then also coupled with misinformation and disinformation indicators; the former refers to false or incorrect information that is spread intentionally or unintentionally (but without realising in both cases that the information is untrue), whereas the latter refers to intentionally false or misleading information that is spread in a calculated way to deceive target audiences. Both mis- and disinformation correspond in essence to disruptive information that misleads and/or misdirects LEAs during their investigations. To this end, TENSOR explores an axiomatic framework based on a combination of theories modelling uncertainty (such as Dempster-Shafer) and machine learning algorithms.

3.2.5 Multimedia Forensics

Multimedia forensics aims to detect digital manipulations (in particular splicing and copy-move manipulations) on online images. The main challenges pertain to the extensive degradation of online content due to the large number of re-savings (between the originally captured image and the image that is published online) and the excessive computational cost of powerful forensic analysis methods. Given the fact that a number of different approaches have been proposed in the literature, each of which has shown to be successful only under specific assumptions and cases (Zampoglou et al., 2017), the TENSOR toolbox implements a number of complementary approaches that can be applied on demand to multimedia content of interest.

3.2.6 Semantic Reasoning and Enrichment

Finally, *semantic reasoning and enrichment* aims to first semantically represent all pertinent information into a network of interconnected ontologies, capitalising on advanced knowledge representation and intelligent context-based reasoning solutions; information from external sources (such as other terrorism-related datasets) can also be integrated into these ontologies. Semantic reasoning is then used to further enrich this data, by deriving facts from the relations between concepts on an individual and collective level so as to enable the detection of unusual event and activity patterns, whilst recognising novel instances of usual patterns.

3.3 MULTI-MODAL SUMMARISATION

In TENSOR, one of the objectives is to present to the users the gist of potentially relevant material discovered on the Web in terms of a summary in the language of preference of the user, and to facilitate the interactive exploration of this material using visual analytics techniques. The summarisation module is still at an early stage of development: the general architecture has been defined, but not all submodules have been integrated or implemented. In this section, we describe the modules needed in order to produce the summaries and visualise the results.

Nowadays, the most popular summarisation strategy is "extractive", which tends to select entire sentences from the original text source(s), based on some relevance metrics. The most relevant sentences are concatenated into a summary; see, e.g., (Diligenti et al., 2004) for an overview. Although extractive summarisation can be realised with little linguistic analysis and the resulting summaries are always grammatically correct, they often lack coherence. Furthermore, the original and the summary are in the same language.

Opposed to extractive summarisation is abstractive summarisation. Starting from a conceptual representation of the original information obtained by language analysis, abstractive summarisation creates intermediate linguistic or conceptual structures from this representation,

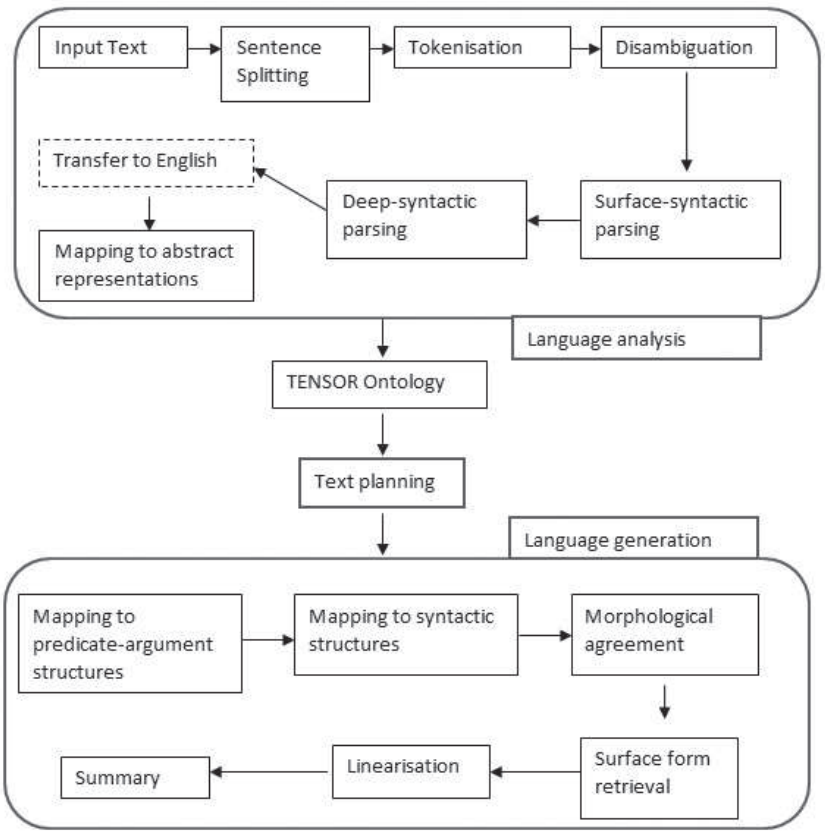
selects the most relevant content chunks and then generates a summary using Natural Language Generation (NLG) techniques. The implementation of our abstractive summariser is based on a sequence of modules that realise the sequence of transitions between different strata. The pipeline can be divided into three main parts:

1. **Language analysis**: Language analysis is carried out by a text analysis pipeline that takes as input the textual content of a document in a given language. This content is first analysed and represented as a forest of abstract syntactic trees. In the case that the input language is different from English, every lexeme in the tree is mapped onto an English lexeme using bilingual dictionaries in order to arrive at a kind of inter-lingua structure that facilitates language neutral representations. These English "inter-lingua" structures are mapped onto semantic structures modelled as RDF triples and then to an ontology.
2. **Text planning**: Conceptual summarisation is approached by assessing the relevance of the semantic structures produced by the language analysis step. Relevance is assessed according to multiple criteria, such as the frequency and joint mention of specific contents in the analysed texts, and lexical-semantic and conceptual relatedness of contents according to lexical databases, sense embeddings and ontologies. Additionally, any inferred knowledge relevant to the domain, the use cases or the production of natural language should also be considered. By considering aspects related to the end user of the system, summaries can be generated tailored to specific users. In addition to determining the relevance of contents, our text planning component also attempts to guarantee a degree of coherence in the summary generated by sorting relevant contents in a sequence that satisfies certain coherence constraints, e.g. grouping together in the text contents that are conceptually related.
3. **Natural language generation**: Following this planning step, linguistic generation starts by transferring the lexemes associated to the semantic structures to the desired target language, using available multilingual lexical resources. Then, the structure of the sentence is determined and all grammatical words are introduced and linked with syntactic relations. Finally, all morphological

agreements between the words are resolved, the words are ordered and punctuation signs are introduced.

In the following, we give more details about the aforementioned modules: language analysis (Sections 3.3.1 to3.3.5), text planning (Section3.3.6), and language generation (Sections 3.3.7 to 3.3.11), as well as visual analytics (Section 3.3.12). Figure 2 shows all the components analysed in the following sections and the connections between them.

FIGURE 2: OVERVIEW OF THE SUMMARISATION PIPELINE



3.3.1 Sentence splitting and tokeniser

Language analysis starts by determining sentence and token boundaries. Rather than addressing tokenisation at a word level, our analysis pipeline treats each sequence of words referring to a specific entity as an atomic unit of meaning. In doing so, we seek to avoid unnecessary internal analysis of multiword expressions which may not even have a strictly compositional meaning (as, e.g., United States of America), and also to eventually obtain predicate-argument structures in which the arguments are not just words, but expressions with an atomic meaning.

3.3.2 Surface-syntactic parsing

In order to determine the syntactic structure of each sentence, we use Bohnet & Nivre's (2012) joint lemmatiser, part of speech tagger, morphology tagger, and dependency parser trained on the CoNLL'09 Penn Treebank dataset (Hajič et al., 2009). This system was the first one to be able to parse non-projective dependency trees while predicting at the same time the part of speech (PoS) and the dependencies, instead of predicting first the PoS and using it for predicting the dependencies in a second step. The authors report an Unlabeled Attachment Score of 93.67, a Labeled Attachment Score of 92.68, and a PoS tagging accuracy of 97.42 (the best possible score being 100 in all cases) in English, improving the state-of-the-art in several languages at the time of publication, and still competitive with current state-of-the-art systems. The sentence splitting, tokenisation and parsing steps require together an average of 65 milliseconds of processing time per sentence.

3.3.3 Deep-syntactic parsing

The objective of this component is to identify and remove all functional words (auxiliaries, determiners, void prepositions and conjunctions) in the surface-syntactic tree and to generalise the syntactic dependencies obtained during the previous stage, while adding sub-categorisation information for lexical predicates. The resulting structures after this step are deep-syntactic trees, in the sense of the Meaning-Text Theory (Mel'čuk, 1988), which is the theoretical framework that underlies the

whole natural language processing pipeline. The mapping between surface and deep syntactic trees can be achieved using rule-based (Mille et al., 2017b) or statistical (Ballesteros et al., 2014) graph-transduction systems. Both systems are able to perform the removal of functional words (*hypernode identification*) with an accuracy of about 99%, and derive the deep dependencies with a recall of about 91% (LAS) in English, for which it is possible to rely on good quality lexical resources (see next section). The deep-syntactic parsing step is currently performed in an average of 25 milliseconds per sentence.

3.3.4 Coreference resolution, word sense disambiguation and entity linking

This step comprises tasks aimed at determining the lexical sense, conceptual meaning or denoted entity of specific words or groups of words in the text. Several state-of-the-art methods and resources for coreference resolution, word sense disambiguation, named entity recognition and entity linking are being considered. Lexical databases and knowledge bases like WordNet (Miller, 1995), PropBank (Kingsbury & Palmer, 2002), VerbNet (Schuler, 2005), FrameNet (Baker et al., 1998), DBPedia (Auer et al., 2007) and BabelNet (Navigli & Ponzetto, 2012) can be used as repositories of senses and entities, possibly extended with domain specific knowledge compiled in collaboration with user partners. For the coreference resolution task, we will experiment with both simple baseline methods, e.g., best mention method based on well-studied syntactic and lexical constraints, and advanced methods such as those implemented in the Standford CoreNLP tools (Manning et al., 2014). Similarly, we will consider a range of methods and tools for the disambiguation and linking tasks, ranging from baselines known to perform well, e.g., most frequent sense, to more complex methods, i.e., those based on features extracted from the local context of mentions of entities, or graph-based global disambiguation methods that aim at producing coherent sets of sense assignments.

3.3.5 Mapping to abstract representations

This component outputs representations that facilitate the mapping to the TENSOR ontologies. For mapping deep-syntactic structures to more abstract linguistic representations, large-scale lexical resources are needed. Unfortunately, such resources are available, at this point, only for English. For this reason, we need to map all input languages to English. Using multilingual resources such as BabelNet, it is possible to obtain the translations of all words into English. Once this is done, the sub-categorisation information in the deep-syntactic structure allows us to obtain Frame annotations on top of connected predicate-argument structures. During this step, shared argumental positions are made explicit and idiosyncratic structuring such as the representation of raising and control verbs is generalised. From the implementation perspective, this task is very similar to that of deep-syntactic parsing, i.e., we are developing graph transducers in order to achieve it. The whole analysis pipeline – from text to abstract representations – has undergone preliminary evaluation in English and obtained Unlabeled Attachment Scores of 74% and 71% for precision and recall respectively (see Mille et al., 2017b), and needs about 150 milliseconds per sentence.

3.3.6 Text planning

Our approach to text planning assumes either a deep linguistic representation with semantic annotations (i.e., disambiguated word senses, links to denoted entities) or a fully conceptual representation based on domain ontologies and upper models if and when it becomes available. As explained before, the main tasks of this module are to assess the relevance of the contents and to structure them in a way that guarantees a coherent presentation in the text. Drawing from the literature in text-to-text summarisation and data-to-text planning, we will experiment with graph-based methods to explore and rank the contents in the semantic repository according to multiple criteria. This method will be supported by recently published resources like semantically annotated corpora and distributional sense embeddings. Additionally, pattern extraction methods will be considered to obtain maximally relevant subsets of contents from the semantic repository, while seeking to ensure that grammatically

and semantically correct clauses and sentences can be generated out of them.

3.3.7 Mapping to output language predicate-argument structures

Starting from the structures provided by the text planning module, first, some idiosyncratic transformations are made to adjust the structures to the predicate-argument format understood by our generation pipeline, and then, the English labels of the nodes are translated into the desired target language using lexicons. These lexicons must not only contain language-specific vocabulary, but also be linked to our pivot language, namely English. Given that BabelNet senses annotated during the analysis stage are language-independent, we will use them as the cross-linguistic link.

3.3.8 Mapping to syntactic structures

Once genuine predicate-argument structures in the target language are available, the first task is to find which node in each structure is most likely to be the root of the dependency tree. That is, we want to identify what will be the main verb of the sentence, or the word that triggers its appearance. Around the main node, the deep-syntacticisation module builds the rest of the syntactic structure of the sentence. In particular, it is able to decide if a main predicate has to be introduced, or what will be realised as an argument, an attribute, or a coordination. The next step in the procedure is to obtain surface-syntactic structures, i.e., to generate all functional words and label the dependencies with language-specific relations, that is, the opposite actions to the ones performed during the deep-syntactic analysis step. As a generator, we also use graph transducers, as described in (Mille et al., 2017a), together with language-specific lexical resources; see, e.g., (Mille & Wanner, 2015). The resulting structure contains all the words that will appear in the final sentence, together with morpho-syntactic features and syntactic dependencies such as *subject*, *object,* etc. that link the words with one another.

3.3.9 Morphological agreement resolution

During the generation of syntactic structures, morphological features of individual words are already inserted (e.g., nominative case for a German subject). During the transition to the morphological structure, agreement is established using the introduced morphological features and the fine-grained syntactic relations in the surface-syntactic structure. For instance, a verb will get its number and person from the element linked to it with the *subject* dependency relation.

3.3.10 Surface form retrieval

The surface forms of the words are retrieved using a full-form dictionary. In order to obtain the full-form dictionary, we will run the morphological tagger of our surface syntactic parser on a large collection of texts and store each possible combination of surface form, lemma and morphological features. We will therefore be able to retrieve a surface form given a lemma and a set of morphological features. The size of the text collection is crucial in order to ensure a large coverage.

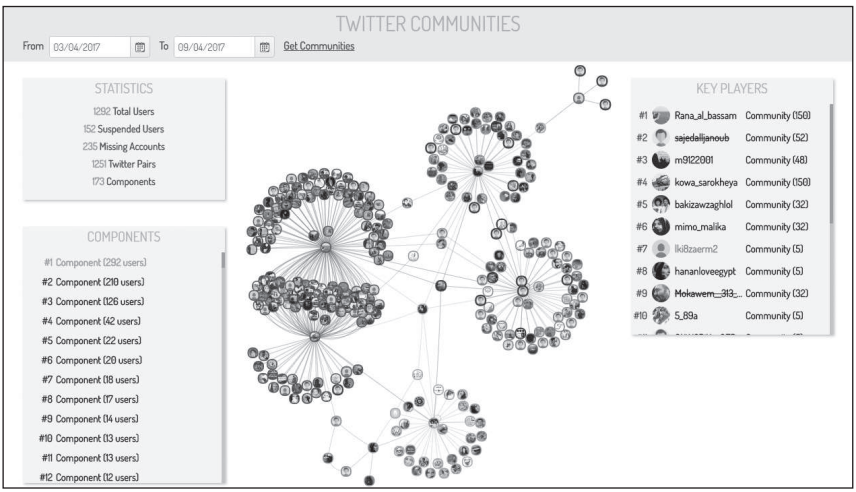
3.3.11 Linearisation

This component takes as input surface-syntactic trees and determines the word order for each tree. In order to ensure large coverage, linearisers will be trained on existing surface-syntactic treebanks, following what has been done in, e.g., (Bohnet et al., 2011), in which the system performs a beam search for optimal word ordering. Words are eventually ordered with a bottom-up method, starting from within subtrees and then ordering subtrees with one another. This linearisation system has obtained high scores on English datasets, with a BLEU score above 89%, and about 58% of the sentences in which all words are ordered exactly as expected.

3.3.12 Visual analytics

Visual analytics algorithms aim to present to the users the analysed and correlated data in a clear and concise interface that allows them to navigate through this information naturally and to improve their situational awareness. For instance, in order to track and flag terrorism-related activities in social media networks, LEAs face big challenges in monitoring relationships and communication activities taking place in such complex networks. To this end, TENSOR proposes a visualisation tool (Andreadis et al., 2017) that offers two novel functionalities based on the key-player identification and key-community detection methods (described in Section 3.2). Both are exposed as a combined Web service and are performed on data from social networks that can be distinguished into statistics, key players and key communities, as demonstrated in Figure 3 for an example from Twitter.

FIGURE 3: SOCIAL NETWORK ANALYTICS VISUALISATION



The network constitutes a straightforward visual representation of how Twitter accounts mention each other and the communities they formulate. Every node in the graph represents a user (profile picture is shown on the node), while every edge is a connection between two users. Communities are indicated by different coloured borders around the

nodes; if an account is inactive, the respective node is coloured red and labelled as "Suspended!" or is coloured black and labelled as "Does not exist!" depending on the case. By clicking on a node, a window pops up to provide more information about the selected user. The pop-up window contains a profile picture on the top left and some account details on the top right, followed by a list of all tweets posted by the featured user. The account details include a name, a username, a description written by the user, a link to the original Twitter page and a label to inform whether the user is suspended or non-existent. Regarding the list of tweets, each item has external links, if any, they are sorted by date and linked to the original tweet. The implementation is based entirely on open-source tools and can be adapted beyond Twitter to instant messaging and other platforms.

3.4 SYSTEM DEVELOPMENT AND IMPLEMENTATION

TENSOR aims to capture, understand and accommodate as many end-user requirements and technical considerations from the earliest stage possible. This section addresses just some of these points by presenting both a high-level overview of the TENSOR architecture and implementation, whilst introducing a small subset of lower-level considerations and challenges where they help elucidate the TENSOR solution.

Developing and implementing a platform such as TENSOR is ambitious and presents many challenges. Although many of these challenges are focussed heavily on the research and delivery of beyond-state-of-the-art tools and technologies aimed to assist LEAs in identifying, monitoring and ultimately combatting terrorism. Behind the scenes there are a myriad of considerations that need to be made to protect the integrity, security, and operation of such tools as well as ensuring straightforward integration and availability. If LEAs are ever going to trust and rely on such technologies, whether independently or within a consolidated platform such as TENSOR, then it is important that under the hood, it is built on an opaque foundation of good pragmatic standards (some of which are introduced in the following sections) and technologies that support the legal and ethical requirements and constraints imposed on LEAs. Furthermore, although TENSOR's output will only be a research

prototype, it is useful to keep in mind some of the proposed security standards required by law enforcement organisations to run an operational product (e.g., PoliceICT, 2017).

3.4.1 TENSOR Architectural Realisation

The TENSOR platform's design loosely follows a Service-Oriented Architecture (SOA) with the main deviation coming from not all components being entirely stateless and the need for a centralised taxonomy and ontology. From a high-level, the system is broken down into many modules or components (defined as services) within three distinctive phases:

1. Discovery and acquisition
2. Analysis and storage
3. Reporting and visualisation

The discovery and acquisition of terrorism-related content will be the source of all data within the system. From here, identified content from a growing collection of *crawl points* will begin its life within the system. It will be given an identity and verifiable hash (see Section 3.4.4) which will uniquely identify each individual piece of content for as long as it is maintained within the system, its archive, or it is deemed irrelevant or unnecessary and is destroyed. This portion of the system consists of several components as discussed earlier in this paper whose responsibility it is to search, crawl, scrape, and interact with various Web- and darknet-based services and sources. The aim is to identify and capture content related to the many aspects of terrorism and terrorist organisations. Once ingested, the data will be passed onto the next phase in order to determine its relevance and validity before it is stored or destroyed.

The analysis and storage phase aims initially to identify for every new piece of content - artefacts and entities - whether it is relevant to the subject of terrorism and whether it is in the interest of both public security and the LEA. TENSOR also aims to capture the key principles of privacy-by-design (Langheinrich, 2001; Cavoukian, 2011) in this phase

to ensure that only data that is absolutely necessary is kept. Any data identified as irrelevant will be immediately destroyed with no further processing, whilst the remainder will be stored in the central content repository. From this point on, select content attributes will be made available on a component-by-component basis for further processing, so that new knowledge and insights can be attained. This higher-level new knowledge will be stored in the central repository and made available to the next phase.

The final phase involves the interaction between the end users and the relevant TENSOR content and knowledge. These components aim to provide the tools for end users to explore and visualise the outputs of the various novel techniques and technologies being developed in TENSOR. Not only that, but they will also have the ability to influence all three phases of the architecture on a case-by-case basis by retrieving and visualising specific sources or types of content and re-configuring the platform's operational focus. This is considered the *value* phase of the system, where the culmination of the TENSOR components and the overall SOA approach provides the end user with insights that they wouldn't have had otherwise.

Across each of these phases will be the growing burden of potentially massive quantities of data that the platform may have to deal with, for which there is little solution other than allocating and managing large storage capacities. For that, clever data management techniques will need to be assessed and implemented. One such issue is the data growth rate, which cannot be mitigated effectively for multimedia given that Web-based multimedia content is often already compressed close to practical limits and although further compression is a possibility, it offers little or no gain. The only real defence is to ensure that multimedia content is not duplicated, achievable by indexing the content hashes (see Section 3.4.4). Textual content on the other hand can be easily, efficiently and effectively compressed automatically via many database technologies.

3.4.2 TENSOR Services Orchestration

For these phases to work effectively, the inter-component communication will be standardised where possible and appropriate. The use of

representational state transfer (REST) interfaces over secure hypertext transfer protocol (HTTPS) connections with JSON and/or XML as the representations of in-transit data appears to have a monopoly in the software industry today. TENSOR makes no aim to deviate from this secure, straight-forward, and trusted approach, but does aim to ensure the best compromises are made between security, complexity and usability are maintained throughout.

At the core of the system will be the central content repository. Not only is this component responsible for managing the storage of all TENSOR content and knowledge, but also the auditing of all activities and interactions with the system. Such auditing is crucial to ensuring the required level of trust and reliability for the use of valuable investigative outcomes in the chain of custody. All of this begins with the fundamental requirement of identity.

3.4.3 Managing Identity

The importance of identity means that every individual piece of content and knowledge created, discovered, stored, or accessed using the TENSOR platform needs to be uniquely identifiable throughout its entire lifetime. TENSOR aims to achieve this using Universally Unique Identifiers (UUIDs) for each and every artefact, entity, and relationship as well as every audited event. Using UUIDs ensures that wherever a piece of data starts its life within the TENSOR platform, it can be allocated an identity without any central coordination whilst confident in the knowledge that the possibility of the same UUID being generated elsewhere in the platform is almost zero.

There are four UUID versions available (Leach et al., 2005), however the expected data growth rates for the platform do not even come close to the operational limitations of properly generated UUIDs, so because of this, they can be chosen based on additional available features and popularity. Version 1 and 4 are the most popular implementations, which can be respectively classified as machine-and-clock based, and securely random. Version 1 is more attractive based on the additional meta-data contained within the UUID. That is a combination of the MAC address of the computer on which it was generated and the time to the nearest

100-nanoseconds, including the central processing unit's (CPU) current clock cycle, making it near impossible to see the same UUID twice over an inconceivable period of time. The additional benefit of version 1 on top of maintaining an audit of the time in which it was generated, is the MAC address which reveals the actual machine that created it. This could be useful for the future verification of the data's original source.

Once all data can be uniquely identified within the platform, it is then possible to log every state change and activity which takes place against each unique piece of data. TENSOR will aim to ensure that this logging is duplicated automatically from the point of creation on both the machine performing the action, in the way of log files, and within the central storage repository in order to ensure multiple sources of the same truth. For every action taking place involving a piece of TENSOR data, its original and unique UUID will be logged with a description of the type of action taking place. Examples of such actions that may be carried out on an artefact or entity include: initial discovery or acquisition; storage; processing; retrieval; visualisation; and removal or archiving.

Each time an action or state change is logged, where there is a process or user responsible, this should also be recorded in order to ensure accountability is always present. TENSOR will aim to ensure that there is no situation where an action can be performed, or the state of the data be changed, where there is no accountable party.

3.4.4 Content Security and Verifiability

Another important and widely used aspect of chain of custody is to ensure the authenticity of the data, primarily to prove that it has not been tampered with in any way since it was originally obtained (Prayudi & Sn, 2015). The current buzz word for LEAs regarding this is "hashing". For example, Interpol already maintains a database of hashed images of child abuse material, which enables the rapid identification of whether a found image is new or not and can help in tracing the source of such images (Interpol, 2017; McCulloch, 2007). A hashing algorithm, as it is known, performs a cryptographic calculation against the underlying data of an artefact, entity, or media file, and generates a relatively short and unique "hash" of the data. The algorithm can easily be re-run at any

point against the subject data and if even 1-bit (or a character in a text file, or a pixel in an image), then the generated hash will be completely different. In fact, this variation in the hashes is what makes them highly reliable. For example, although it may be possible to find or generate two sets of data that would lead to the same hash, it is extremely unlikely that both of these could be valid data of the same type - such as both being images, never mind images that are similar. But, it remains critical that a suitable hashing algorithm, such as the Secure Hashing Algorithm is employed and done so correctly.

TENSOR's aim is to go one step further than hashes alone by investigating the viability and potential additional trust in authenticity gained by combining hashes with the use of Digital Signature Algorithm technology (Kravitz, 1993). Where, for instance, a hash cannot be changed to result in the same media, there is no guarantee that the media hasn't been changed, along with the hash given alongside it. Using DSA would allow the content to be hashed in the same way as before, but to also *sign* the hash with a #verifiable digital signature. It would thus be possible to verify not only that the subject media (or artefact) is authentic, but also that it was actually created by a given component and has not been tampered with.

On the subject of tampering, another important consideration in the development of the TENSOR platform is ensuring the accountability and auditing of activities and that data cannot easily, if at all, be manipulated. Primarily, the project aims to explore write-once read-many (WORM) technologies. These however, can be expensive and require more complex physical hardware and processes to be in place. So, whilst the aim is to investigate and outline recommendations in the use of such technology, softer approaches should also be considered. One way this could be achieved is by implementing tightly secured write-once database restrictions with insert-only privileges and ensuring regular hard backups.

Generally, within the platform, the aim is to mandate effective data security practices across the board, but particularly within the central content repository. Such mandates include, but are not limited to, the use of good standards-based encryption throughout. Primarily this occurs at two obvious points: when data is in transit; and when data is at rest.

Transport Layer Security (TLS) is widely used and heavily standardised. It is also relatively easy to use and configure and should be employed for all inter-component or inter-server communications, even within internal systems. Encryption at rest, on the other hand, deals with any data stored on a physical or virtual storage device, whether live or backup. Again, this is relatively easy to implement and is offered by many database management systems with the use of strong standards-based encryption algorithms. It should be noted that ensuring encryption *keys* are securely managed and all core principles around the chosen standards or algorithms is adhered to is vital for success.

3.4.5 Future Proofing Architecture

Containerisation of software applications, particularly in the micro-services variation of SOA, has been a rapidly growing area of development in the industry. It allows a software application to be wrapped in a standardised container, along with its production configuration and its very own operating system stack - including firewalls. These containers can be rapidly deployed in a standard way and often need only to be plugged together.

TENSOR aims to utilise containerisation, in particular using Docker (https://www.docker.com/), to enable the development teams of each component to configure and deploy how they see fit following black-box principles. As long as each partner respects the agreed inter-component communication protocols, then much of the production-level integration becomes effectively a *plumbing* problem. On top of this, the use of scalable messaging platforms within the platform integration layer is also being considered, which aims to exploit powerful high-availability software principles.

Finally, it is envisaged that TENSOR will not only achieve many exciting and challenging research goals with its many state-of-the-art and beyond-state-of-the-art tools and techniques, but it will also be built in such a way as to enable the highest possible technology readiness and best possible future exploitation opportunities. Much of this will be a result of the project's forward-thinking architectural focus and effective partner cooperation.

4. LEGAL AND ETHICAL ASSESSMENT

The improvement of current regulatory framework is one of the main scopes of TENSOR. A harmonised legal framework is of foremost importance when it comes to the cross-border cooperation of LEAs. In this chapter, we present a general overview of the existing legal procedures in Spain, Greece, Germany and the United Kingdom related to the crime of terrorism.

4.1. SPAIN

In Spain, there are five different public organisations (in three different levels: National, State and Regional) with authorisation to deal with counter terrorism affairs.

The Spanish Criminal Code provides special penal sanctions against terrorism by punishing those who belong to, serve, or collaborate with terrorist organisations or groups (Ministerio de Justicia-Secretaria General Tecnica, 2013). The key approach was the definition of a terrorist organisation or group and the classification of all related illegal behaviours, such as participation in terrorist organisations or groups, and/or simple collaboration with them. Furthermore, the Criminal Code also considers as crimes, individual terrorism and other new types of conduct which impact on the international community, including computer criminal offences.

There are some provisions in the Spanish Criminal Procedure Law (Ministerio de Gracia y Justicia, 2015) regulating the interception of telephone and telematic communications. However, a judicial authorisation has to be issued in order to legally use these investigative methods. The competent Magistrate or the Public Prosecution Services may authorise the Judiciary Police to act under an assumed identity in undercover operations. The undercover agent can only carry out actions necessary to the investigation and proportionate with its purpose. The Spanish Organic Law 15/1999 (BOE-A-1999-23750,1999) intends to guarantee and protect the public liberties and fundamental rights of people regarding the

processing of their personal data. However, this Law and the General Data Protection Regulation are not applicable to the processing of files related to the investigation of terrorism and serious forms of organised crime, or the investigation of other serious criminal offences.

Furthermore, the current Spanish legislation does not provide any reference for the use of search robots in police work. The Criminal Procedure Law introduces provisions concerning the retention of data and other information contained in computers or other electronic devices in order to preserve the integrity and eligibility of these materials in court proceedings.

4.2. GREECE

An electronic investigation should always respect the fundamental human rights stipulated in articles 19 of the Constitution (Hellenic Parliament, 2008) on secrecy of letters and all other forms of free correspondence or communication and 9A of the Constitution for the protection of personal data. However, the Greek legal framework recognises exceptions to the absolute character of these rights for reasons of national security or for the investigation of especially serious crimes, under articles 3 and 4 of Law 2225/1994 (Hellenic Parliament, 1994) on the confidentiality of communications, and article 253A of the Criminal Procedure Code (Hellenic Parliament, 2004) on the investigation of criminal groups. Within this scope, the Hellenic Police can proceed to an undercover investigation, after formal authorisation has been issued by the Prosecutor, or the Prosecutors' Council.

Generally, there are no restrictions on collecting evidence for judicial purposes or police investigations in the case of a serious offence under article 251 of the Criminal Procedure Code (Hellenic Parliament, 2003). The admissibility of automatically generated evidence should also be considered, meaning that as acceptable evidence in court is considered the electronic evidence could be recreated. The Greek legal system stipulates in article 46 paragraph 2 of the Penal Code (Hellenic Parliament, 1951) that whoever intentionally incites others to commit a crime is considered as an agent provocateur. Only the participation in a pre-planned

illegal act, within the framework of an official judicial order could be exonerated.

4.3. GERMANY

In Germany, the term terrorist offence describes every act that aims to seriously intimidate the population or to force or deter public authorities or international organisations from doing something. In addition, terrorist offence describes the act of destabilising or destroying the political, constitutional, economic, or social basic structures of Germany or an international organisation. The lawful interception is regulating the monitoring of telecommunications activities and contents. The legal basis is given by the respective laws such as *§100a of the Criminal Procedure Code (Code of Criminal Procedure*, 2014), the G10 Commission (Basic Law for the Federal Republic of Germany, 2014) and *§23a of the Cu*stoms Investigation Service Act (Germany, 2013). German Intelligence Services as well as LEAs could work undercover to obtain information to prevent and detect crime or disorder and maintain public safety.

In Germany, it is a fundamental right to ensure the confidentiality and the integrity of information technology systems, in order to protect the personal data stored or processed in information technology systems. Infringements of this right are possible within narrow bounds. Preventive state interventions – especially in the framework of online searches – are only permissible constitutionally, if factual indications exist of a concrete danger to a predominantly important legal interest.

The usage of search robots is not mentioned within the German legal framework. However, the collection of special types of personal data is permissible only in so far as inter alia such collection concerns data which the data subject has evidently made public or such collection is necessary in order to avert a substantial threat to public safety. The collection and storage of terrorist content for the purpose of the evaluation of evidence, danger prevention, or the prosecution of a criminal offence is not illegal in Germany. The act of encouraging an individual to commit a crime violates the basic principle of fair proceedings.

4.4. UNITED KINGDOM

First and foremost, terrorism is a crime, which has serious consequences and requires to be distinguished from other types of crime. Individuals who commit terrorism-related offences contrary to UK law are subject to the processes of the Criminal Justice System and those who are otherwise believed to be involved in terrorism are subject to restrictive executive actions.

The British LEAs use all available powers and tactics to prevent and detect crime or disorder and maintain public safety. Undercover policing is one of those tactics. Applied correctly, and supported by appropriate training, it is a proportionate, lawful, and ethical tactic which provides an effective means of obtaining evidence and intelligence, and includes the identification of online terrorist content. The purpose of undercover police officers is to detect or prevent a more serious crime, and to allow an undercover asset to gain the trust of the criminals they are trying to infiltrate. English law offers a defense to someone accused of a crime if they can show an officer acted as an agent provocateur, i.e., they initiated or instigated the crime.

The Data Protection Act 1998 (DPA) (Great Britain, 1998) is the primary piece of UK legislation governing the protection of data. At the heart of the DPA is a set of eight principles, which deal with the collection, use, quality, and security of personal data and with data subjects' rights.

Public authorities can use online research and investigation tools for a specific and legitimate objective – such as preventing or detecting crime, proportionate to the objective in question and in accordance with the law – but they must ensure not to interfere with a person's right to privacy.

The collection of online illegal content by UK LEAs is governed by the Regulation of Investigatory Powers Act 2000 (RIPA) (Great Britain, 2000), regulating the powers of public bodies to carry out surveillance and investigation, and covering the interception of communications. It is an offence to intercept post/public telecommunications within the UK unless authorised under RIPA or another statute (or have consent). A national best practice guide for Digital Evidence has been produced to provide guidance not only to law enforcement, but all stakeholders who assist in investigating cyber security incidents and crime (Williams, 2012).

5. IMPACT

The TENSOR research and designed prototype solutions will have a significant impact on several security operational challenges. The social impact of deploying the TENSOR solution in operational environments will enable LEAs and Security Agencies to increase accuracy towards actionable threat intelligence, make more informed decisions and deliver elevated preventive power. Delivering the platform in the LEAs operational settings will contribute to increased public safety and reduced risk of terrorist activities, whilst protecting fundamental human rights, such as freedom of expression and privacy, thanks to the built-in data protection and anonymisation capabilities of the platform. The early warning of terrorist content or the emergence of networks will allow for early interventions, allowing prevention of radicalisation without criminalisation of subjects.

TENSOR will also contribute to technical and scientific fields. Its innovation activities will improve Web crawling techniques for faster, more efficient content detection and gathering. Research will also focus on effective content gathering from hard to reach silos on the Dark Web and will deliver better information extraction techniques that can deal with larger amounts of multimedia and multilingual content, enable the processing of highly diverse and previously under-utilised online content. Finally, it will improve automated analysis and data mining approaches that help identify relationships between content, the identification of narratives and trends, and the extraction of spatio-temporal patterns of interest.

5.1 IMPACTS ON HOW LEAS FIGHT TERRORISM

TENSOR will provide a unified platform that enables LEAs to effectively detect, categorise, analyse, reason over, and summarise terrorist-generated content. Ultimately, this will increase LEAs capabilities in detecting and preventing terrorist activities organised via the Web, culminating in increased security and resilience across the EU. TENSOR will empower

LEAs to scale their responsiveness and effectiveness through the horizontal diffusion of information. It will also ensure LEAs benefit from a greater range of operational responses thanks to the early identification of terrorist generated content.

The platform will also leverage intelligent mechanisms that identify potential emerging terrorist activities planned and organised via the Internet and make use of enhanced capabilities to support the early detection and identification of online radicalisation.

It is envisaged that the research will also support the deployment of more effective techniques for distinguishing non-harming religious (or other) extremist ideologies from violent radicalisation activities and employ more effective capabilities in gathering data from the Dark Web, which were previously hidden or inaccessible to them. The solutions will also identify patterns as well as uniform responses and prevention measures, which will be undertaken at a strategic level. These impacts are essential to the operational delivery of counter terrorism security in today's ever-changing world.

5.2 ECONOMIC IMPACT

Open Source Intelligence (OSINT) campaigns for law enforcement and counter-terrorism work have become a complex and resource intensive task for both Government and Defence intelligence agencies. OSINT work has gained momentum to become recognised as a legitimate area of intelligence operations, alongside the more traditional intelligence domains such as HUMINT (agent handling) and SIGINT (signals intelligence). This is particularly true in the domain of counter-terrorism. Nearly all Government and Defence intelligence agencies have resources dedicated to the production of OSINT within the intelligence cycle in order to meet their intelligence requirements and to produce actionable outputs.

The security and ICT market segments that are directly addressed by the TENSOR technologies amount globally to approximately 100B USD and one million jobs with conservative estimates. Supporting the

development of TENSOR will result in a highly novel and competitive platform, and an accompanying ecosystem of companies (large ICT providers and SMEs that are part of the consortium, but also companies that are early adopters of the TENSOR technology). This will help European companies that are active in this market segment increase their market share and achieve higher growth rates. Accordingly, we foresee a proportional increase in the number of jobs related to the TENSOR ecosystem (technology development, training, support, sales, etc.). For TENSOR to be able to affect 1% of the pertinent global market will mean to capture 1 billion USD value and to create (or sustain) 10 thousand related jobs. Given the increasing trends of the market and the growing importance of the security sector, TENSOR is on track to deliver a significant and sustainable impact on the European economy.

6. CONCLUSIONS

The internet provides a haven for the creation, sharing, and access to terrorism-related content. It can be a breeding ground for radicalisation and violent extremism, and is one that is largely going unchecked due to the difficulties LEAs have in accessing, analysing and then managing such large amounts of information. The TENSOR platform will, through efficient data capture, text and multimedia processing, analysis and visualisation of such terrorist content, be able to reduce the workload on intelligence analysts and provide operational benefits in the linking of intelligence extracted from such content. TENSOR will serve the operational requirements of LEAs today and in the future by utilising state-of-the-art technologies and algorithms, and by collaborating closely with a number of LEAs that operate on the frontline of Europe's effort to counter the spread of terrorism and violent extremism.

ABBREVIATIONS

CoA: Course of Action
CNNs: Convolutional Neural Networks
CPU: Central processing unit
DPA: Data Protection Act
DSA: Digital Signature Algorithm
FCA: Formal Concept Analysis
HSMs: Hidden Service Marketplaces
HTTPS: Interfaces over secure hypertext transfer protocol
HUMINT: Human Intelligence - Agent handling
IP2: Invisible Internet Project
LEAs: Law Enforcement Agencies
MEB: Mapping Entropy Betweenness
NLG: Natural Language Generation
OSINT: Open Source Intelligence
REST: Representational state transfer
RIPA: Regulation of Investigatory Powers Act
RNNs: Recurrent Neural Networks
SHA: Secure Hashing Algorithm
SIGINT: Signals intelligence
SOA: Service-Oriented Architecture
TLS: Transport Layer Security
TOR: The Onion Router
UUIDs: Universally Unique Identifiers
WORM: Write-once read-many

Contacts:

Babak Akhgar
CENTRIC
Sheffield Hallam University, UK
E-mail: b.akhgar@shu.ac.uk

Pierre Bertrand
Thales Group, La Défense, France
E-mail:
pierre.bertrand@thalesgroup.com

Christina Chalanouli
KEMEA
Leof. Mesogeion 96
Athina 115 27, Greece
E-mail:
c.chalanouli@kemea-research.gr

Tony Day
CENTRIC
Sheffield Hallam University, UK
E-mail: t.day@shu.ac.uk

Helen Gibson
CENTRIC
Sheffield Hallam University, UK
E-mail: h.gibson@shu.ac.uk

Dimitrios Kavallieros
KEMEA
Leof. Mesogeion 96
Athina 115 27, Greece
E-mail:
d.kavallieros@kemea-research.gr

Emmanuel Kermitsis
KEMEA
Leof. Mesogeion 96
Athina 115 27, Greece
E-mail:
m.kermitsis@kemea-research.gr

Ioannis Kompatsiaris
Information Technologies Institute
Centre for Research and Technology Hellas
6th Klm Charilaou-Thermi Rd
Thessaloniki, 57001, Greece
E-mail: ikom@iti.gr

Eva Kyriakou
European Organisation for Security
Rue Montoyer 10, Brussels
Belgium
E-mail:
Eva.kyriakou@eos-eu.com

George Leventakis
KEMEA
Leof. Mesogeion 96
Athina 115 27, Greece
E-mail: gleventakis@kemea.gr

Euthimios Lissaris
KEMEA
Leof. Mesogeion 96
Athina 115 27, Greece
E-mail:
e.lissaris@kemea-research.gr

Simon Mille
Universitat Pompeu Fabra, Spain
E-mail: simon.mille@upf.edu

Dimitrios Myttas
KEMEA
Leof. Mesogeion 96
Athina 115 27, Greece
E-mail:
d.myttas@kemea-research.gr

Theodora Tsikrika
Information Technologies
Institute
Centre for Research and
Technology Hellas
6th Klm Charilaou-Thermi Rd
Thessaloniki, 57001, Greece
E-mail:
theodora.tsikrika@iti.gr

Stefanos Vrochidis
Information Technologies
Institute
Centre for Research and
Technology Hellas
6th Klm Charilaou-Thermi Rd
Thessaloniki, 57001, Greece
E-mail: stefanos@iti.gr

Una Williamson
Police Service of Northern Ireland
E-mail:
Una.Williamson@psni.pnn.police.uk

REFERENCES AND SOURCES

Andreadis, S., Gialampoukidis, I., Kalpakis, G., Tsikrika, T., Papadopoulos, S., Vrochidis, S., & Kompatsiaris, I. (2017). "A Monitoring Tool for Terrorism-related Key-players and Key-communities in Social Media Networks." In Proceedings of the IEEE European Intelligence and Security Informatics Conference (EISIC 2017), p. 166.

Andrews, S. (2011). "In-close2, a high performance formal concept miner." In Conceptual Structures for Discovering Knowledge, Proceedings of the 19th International Conference on Conceptual Structures (ICCS 2011), pp.50-62.

Association of the Chiefs of Police Officers (2013). "Online research and investigation." College of Policing. Available at: http://library.college.police.uk/docs/appref/online-research-and-investigation-guidance.pdf [Accessed 21 Aug. 2017]

Auer, S., Bizer, C., Kobilarov, G., Lehmann, J., Cyganiak, R., & Ives, Z. (2007). "DBpedia: A nucleus for a Web of open data." The Semantic Web, pp. 722-735.

Baker, C. F., Fillmore, C. J., & Lowe, J. B. (1998). "The Berkeley FrameNet project." In Proceedings of the 36th Annual Meeting of the Association for Computational Linguistics (ACL 1998) and 17th International Conference on Computational Linguistics (COLING 1998), Volume 1, pp. 86-90.

Ballesteros, M., Bohnet, B., Mille, S. & Wanner, L. (2014). "Deep-Syntactic Parsing." In Proceedings of the 25th International Conference on Computational Linguistics (COLING 2014), pp. 1402-1413.

Basic Law for the Federal Republic of Germany in the revised version published in the Federal Law Gazette Part III, classification number 100-1, as last amended by Article 1 of the Act of 23 December 2014 (Federal Law Gazette I p. 2438).

Blondel, V. D., Guillaume, J. L., Lambiotte, R., & Lefebvre, E. (2008). "Fast unfolding of communities in large networks." Journal of Statistical Mechanics: Theory and Experiment, vol. 2008, no. 10, p. P10008.

Bohnet, B. & Nivre, J. (2012). "A transition-based system for joint part-of-speech tagging and labeled non-projective dependency parsing." In Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning (EMNLP-CoNLL 2012), pp. 1455–1465.

Bohnet, B., Mille, S., Favre, B., & Wanner, L. (2011). "<StuMaBa>: from deep representation to surface." In Proceedings of the 13th European workshop

on Natural Language Generation (ENLG 2011), Surface-Generation Shared Task, pp. 232-235.

Bouchard, M., Joffres, K., & Frank, R. (2014). "Preliminary analytical considerations in designing a terrorism and extremism online network extractor". In Computational Models of Complex Systems, pp. 171-184.

BOE-A-1999-23750 (1999). "The Data Protection Act Law (Ref. BOE-A-1999-23750)". [online] Available at http://www.boe.es/buscar/doc.php?id=BOE-A-1999-23750 [Accessed 21 Aug. 2017]

Cavoukian, A. (2011). "Privacy by Design The 7 Foundational Principles". Information and Privacy Commissioner of Ontario. [online] Available at: http://www.privacybydesign.ca/ [Accessed 21 Aug. 2017]

Chen, H. (2011). "Dark Web: Exploring and data mining the dark side of the Web," Vol. 30. Springer Science & Business Media.

Clauset, A., Newman, M. E., & Moore, C. (2004). "Finding community structure in very large networks," Physical Review E, vol. 70, no. 6, p. 066111.

Code of Criminal Procedure in the version published on 7 April 1987 (Federal Law Gazette [Bundesgesetzblatt] Part I p. 1074, 1319), as most recently amended by Article 3 of the Act of 23 April 2014 (Federal Law Gazette Part I p. 410).

Danon, L., Diaz-Guilera, A., Duch, J., & Arenas, A. (2005) "Comparing community structure identification," Journal of Statistical Mechanics: Theory and Experiment, vol. 2005, no. 09, p. P09008.

Diligenti, M., Gori, M., & Maggini, M. (2004). "A unified probabilistic framework for eb page scoring systems."" IEEE Transactions on Knowledge and Data Engineering, vol. 16, no. 1, pp. 4–16.

Furche, T., Gottlob, G., Grasso, G., Schallhart, C., & Sellers, A. (2013). "OXPath: A language for scalable data extraction, automation, and crawling on the Deep Web." The VLDB Journal, 22(1), 47-72.

Ganter, B. & Wille, R. (1998). "Formal Concept Analysis: Mathematical Foundations", Springer-Verlag, Berlin.

Germany, Customs Investigation Service Act (Zollfahndungsdienstgesetz), 16 August 2002, last amended 20 June 2013. [online] Available at: http://www.gesetze-im-internet.de/zfdg/BJNR320210002.html [Accessed 21 Aug. 2017]

Gialampoukidis, I., Kalpakis, G., Tsikrika, T., Papadopoulos, S., Vrochidis, S., & Kompatsiaris, I. (2017). "Detection of Terrorism-related Twitter Communities using Centrality Scores." In Proceedings of the 2nd International Workshop on Multimedia Forensics and Security, pp. 21-25.

Gialampoukidis, I., Kalpakis, G., Tsikrika, T., Vrochidis, S., & Kompatsiaris, I. (2016a). "Key player identification in terrorism-related social media

networks using centrality measures." In Proceedings of the IEEE European Intelligence and Security Informatics Conference (EISIC 2017), pp. 112-115.

Gialampoukidis, I., Tsikrika, T., Vrochidis, S., & Kompatsiaris, I. (2016b). "Community Detection in Complex Networks Based on DBSCAN* and a Martingale Process." In Proceedings. of the 11th IEEE International SMAP Workshop, pp. 1-6.

Great Britain (1998), Data Protection Act. London: Stationery Office. [online] Available at: http://www.legislation.gov.uk/ukpga/1998/29/contents [Accessed 21 Aug. 2017].

Great Britain (2000), Regulation of Investigatory Powers Act. London: Stationery Office. [online] Available at: https://www.legislation.gov.uk/ukpga/2000/23/contents [Accessed 21 Aug. 2017].

Hajič, J., Ciaramita, M., Johansson, R., Kawahara, D., Martí, M. A., Màrquez, L., Meyers, A., Nivre, J., Padó, S., Štěpánek, J., Straňák, P., Surdeanu, M., Xue, N., & Zhang, Y. (2009). "The CoNLL-2009 shared task: syntactic and semantic dependencies in multiple languages." In *Proceedings of the Thirteenth Conference on Computational Natural Language Learning: Shared Task* (CoNLL 2009), pp. 1-18.

He, Y., Xin, D., Ganti, V., Rajaraman, S., & Shah, N. (2013). "Crawling Deep Web entity pages." In Proceedings of the 6th ACM international conference on Web Search and Data Mining (WSDM 2013), pp. 355-364.

Hellenic Parliament (1951), Penal Code.

Hellenic Parliament (1994), For the protection of free correspondence and communication and other provisions.

Hellenic Parliament (2003), Article 251 of the Criminal Procedure Code.

Hellenic Parliament (2004), Article 253A of the Criminal Procedure Code.

Hellenic Parliament (2008), The Constitution of Greece, as revised by the parliamentary resolution of May 27th 2008 of the VIIIth Revisionary Parliament.

Interpol (2017). "Crimes against children, victim identification." Interpol. [online] Available at: https://www.interpol.int/Crime-areas/Crimes-against-children/Victim-identification [Accessed 21 Aug. 2017].

Kravitz, D. W. (1993). U.S. Patent No. 5,231,668. Digital signature algorithm. Washington, DC: U.S. Patent and Trademark Office.

Lancichinetti, A., Fortunato, S. & Radicchi, F. (2008) "Benchmark graphs for testing community detection algorithms," Physical Review E, vol. 78, no. 4, p. 046110.

Langheinrich, M., (2001). "Privacy by design – principles of privacy-aware ubiquitous systems." In Proceedings of the Third International Conference on Ubiquitous Computing (Ubicomp 2001), pp. 273-291.

Leach, P. J., Mealling, M., & Salz, R. (2005). "A universally unique identifier (UUID) URN namespace". Available at: https://tools.ietf.org/html/rfc4122 [Accessed 21 Aug. 2017].

National Police Chief Council (2015). "NPCC Guidance on Open Source Investigation / Research". Kent and Essex Police. https://www.suffolk.police.uk/sites/suffolk/files/003525-16_npcc_guidance_redacted.pdf [Accessed 1 October 2017]

Kingsbury, P. & Palmer, M. (2002). "From TreeBank to PropBank." In Proceedings of the Third International Conference on Language Resources and Evaluation (LREC 2002), pp. 1989-1993.

Manning, C. D., Surdeanu, M., Bauer, J., Finkel, J. R., Bethard, S., & McClosky, D. (2014). "The Stanford CoreNLP Natural Language Processing Toolkit". In Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (ACL 2014) - System Demonstrations, pp. 55-60.

Mel'čuk, I. (1988). "Dependency Syntax: Theory and Practice." State University of New York Press, Albany.

Mille, S., Carlini, R., Burga, A. & Wanner, L. (2017a). "FORGe at SemEval-2017 Task 9: Deep sentence generation based on a sequence of graph transducers." In Proceedings of the 11th International Workshop on Semantic Evaluation (SemEval-2017), pp. 920-923.

Mille, S., Carlini, R., Latorre, I. & Wanner, L.. (2017b). "UPF at EPE 2017: Transduction-based Deep Analysis." In Proceedings of the 2017 Shared Task on Extrinsic Parser Evaluation (EPE 2017), pp. 80-88.

Mille, S. & Wanner, L. (2015). "Towards large-coverage detailed lexical resources for data-to-text generation." In Proceedings of the first workshop on data to text generation.

Miller, G. A. (1995). "WordNet: a lexical database for English." Communications of the ACM, vol. 38, no. 11, pp. 39-41.

Ministerio de Gracia y Justicia (2015), Criminal Procedure Law (approved by Royal Decree of September 14, 1882, and amended up to Organic Law No. 13/2015 of October 5, 2015) (Ref. BOE-A-1882-6036). Available at: http://www.wipo.int/wipolex/en/details.jsp?id=16706 [Accessed 21 Aug. 2017]

Ministerio de Justicia-Secretaria General Tecnica (2013), Criminal Code.

McCulloch, H. (2007) "International Child Sexual Exploitation Image Database". ICPO Interpol. [online] Available at: http://cf.cdn.unwto.org/sites/all/files/docpdf/21sttaskforcemeetingreport2007novmcculloch.pdf [Accessed 21 Aug. 2017]

Navigli, R. & Ponzetto, S. P. (2012) "BabelNet: The automatic construction, evaluation and application of a wide-coverage multilingual semantic network." Artificial Intelligence, vol. 193 pp. 217-250.

Petkos, G., Schinas, M., Papadopoulos, S., & Kompatsiaris, I. (2017). "Graph-based multimodal clustering for social multimedia." Multimedia Tools and Applications, vol. 76, no. 6, pp. 7897–7919.

PoliceICT (2017). "Security Open Standards." Metropolitan Police Service. Available at: https://ict.police.uk/national-standards/security/security-open-standards/ [Accessed 21 Aug. 2017]

P. Pons & M. Latapy, (2006) "Computing communities in large networks using random walks." Journal of Graph Algorithms and Applications, vol. 10, no. 2, pp. 191–218.

Prayudi, Y., & Sn, A. (2015). "Digital chain of custody: State of the art." International Journal of Computer Applications, vol. 114, no. 5.

Rand, W. M. (1971) "Objective criteria for the evaluation of clustering methods," Journal of the American Statistical Association, vol. 66, no. 336, pp. 846–850.

Schuler, K. K. (2005) "VerbNet: A broad-coverage, comprehensive verb lexicon." Ph.D. Dissertation. University of Pennsylvania, Philadelphia, PA, USA. AAI3179808.

Trottier, D. (2015). "Open source intelligence, social media and law enforcement: Visions, constraints and critiques." European Journal of Cultural Studies, vol. 18, no. 4-5, pp. 530-547.

Williams, J. D. (2012). "ACPO Good Practice Guide for Digital Evidence." Metropolitan Police Service.

Weninger, T., Palacios, R., Crescenzi, V., Gottron, T., & Merialdo, P. (2016). "Web Content Extraction: a MetaAnalysis of its Past and Thoughts on its Future." ACM SIGKDD Explorations Newsletter, vol. 17, no. 2, pp.17-23.

Zampoglou, M., Papadopoulos, S., & Kompatsiaris, I. (2017). "Large-scale evaluation of splicing localization algorithms for Web images." Multimedia Tools & Applications, vol. 76, no. 4, pp. 4801-4834.

Zhao, F., Zhou, J., Nie, C., Huang, H., & Jin, H. (2016). "SmartCrawler: A Two-stage Crawler for Efficiently Harvesting Deep-Web Interfaces." IEEE Transactions on Services Computing, vol. 9, no. 4, pp. 608-620.

Zhou, Y., Qin, J., Lai, G., Chen, H., & Reid, E. (2005). "Building knowledge management system for researching terrorist groups on the Web." In Proceedings of the 11th Americas Conference on Information Systems (AMCIS 2005), pp. 344.